No Killer Robots, OK, But What Else Should AI Promise Not to Do?

Let's not have machines make life or death decisions, shall we?

Kevin McCaney

Fri, 08/03/2018 - 10:23

Illustration: Grandfailure/iStock

More than 2,400 scientists and tech leaders working at the forefront of artificial intelligence stepped up to the plate in July, signing a pledge they would take no part in building killer robots.

At the 2018 International Joint Conference on Artificial Intelligence in Stockholm, industry heavyweights including Elon Musk, the co-founders of Google's DeepMind, and quite a few AI research pioneers agreed they would "neither participate in nor support the development, manufacture, trade, or use of lethal autonomous weapons."

The pledge was put on the table by the Future of life Institute. And while you might be suspecting it won't quite settle the question once and for all, it is a reasonable gesture. After all, machines aren't pulling the triggers yet — and the Defense Department, for one, has strict rules limiting any decision to fire to authorized humans — but drones like the Reaper are the delivery boys for Hellfire missiles, and unmanned vehicles of all stripes are becoming increasingly autonomous. Terminator-style humanoid robots are still science fiction (for now), but it's not such a leap from the operator-controlled machines of this minute to AI-powered systems that can call their own shots, as it were.

But aside from the killer robot question, the pledge raises the critical underlying issue of ethics in AI, and whether we put too much faith in a computerized system just because it can remember everything it ever learned, crunch numbers faster than a million people combined, and maybe talk in a soothing, confident voice. Perhaps the brave leaders of AI shouldn't stop at bullets and bombs, but make a pledge concerning the other ways AI can do damage. Here are a few examples.

First, Do No Harm

Some of AI's most impressive early feats have come in the field of medicine. Some tests show AI to be more accurate than doctors at diagnosing heart disease and lung cancer, as well as other illnesses, while providing a variety of support functions . No systems have made more rounds than IBM's Watson, which has multiple versions tailored for specific fields and has garnered praise for its work in high-speed genomic sequencing and clinical trials in cancer research, among other assignments.

But along with those "ah ha!" moments have come of few unsettling "oh-ohs." One other shoe dropped recently when STAT News reported internal IBM documents showed "multiple examples of unsafe and incorrect treatment recommendations" made two years ago by Watson for Oncology. According to the report, Watson had given inaccurate treatment recommendations that ran counter to guidelines from the National Comprehensive Cancer Network and other national bodies, leading some doctors to tell IBM privately they considered Watson unfit for treating patients.

IBM and Memorial Sloan Kettering Cancer Center, which worked on training Watson, pointed out the examples cited were two years old, likely occurred during system testing, and that IBM has made continual upgrades to Watson since. Also, no patients were harmed in the examples cited. But even if Watson’s mistakes weren’t as bad as they might seem, they still show that AI systems are, like humans, fallible.

They also could be led astray. A Harvard study published in May found deep learning medical systems could be vulnerable to adversarial attacks, which target machine learning systems in an effort to make them make a mistake. In medical cases, it could involve manipulating images (say, in an X-ray or retinal scan), causing a system into incorrectly identify an object, leading to a false diagnosis.

As with DOD’s plans for human-machine teaming, the medical world isn’t looking to just hand over the stethoscope and prescription pad to a machine, but to use AI as a source of information for doctors. But medical professionals might want to be a little cautious about some of the information they receive. Maybe AI systems could promise to get a second opinion.

Ready, Aim, Hire, Fire

AI is a powerful tool for human resources departments, managing applications, evaluating credentials, and in some cases, staying in touch with prospective employees via bot-generated text messages, making the whole process more efficient. What could go wrong?

In addition to potentially being taken out of the running because of your response to a bot’s text, AI systems have shown to reflect “algorithmic biases” that came aboard with their original programming, which tech leaders in Silicon Valley say are difficult to remove. Considering that one of the selling points of AI screening systems is that they could remove personal biases from the hiring process, in some cases it may have only shifted, rather than removed, the problem.

There is also the case of a Los Angeles man who was fired by an HR machine that misread some personnel movements and coldly set him packing. It took his bosses three weeks to find out why he was fired.

And on the subject of jobs, there’s an ongoing debate.

AI surely threatens to take a lot of jobs currently done by people, just like more crude but automated machines did before. However, some projections are pretty optimistic about jobs overall, showing a net gain once AI technology is applied to the job market. A recent study from PwC focusing on the U.K. estimates AI will create about as many jobs as it takes away, although the fields of employment will shift — fewer jobs in transportation and manufacturing, for instance, but more new jobs in healthcare. The PwC report echoes some other assessments such as Gartner’s prediction last year that while AI would eliminate 1.8 million jobs by 2020, it would create another 2.3 million during the same time.

Because AI is so good at predictive analytics, maybe it should promise to figure out what people losing their jobs should do next before sending them packing.

And once AI gets solid on weapons control, curing cancer and employment, maybe it could make some other promises, such as not recommending bad movies, or bad dates, or ordering dollhouses when an Echo device overhears a TV news report.

Ultimately, of course, it’s not up to the machines. It’s up to how much power humans decide to give them.

View printer friendly version

Artificial Intelligence
robots
Federal Cybersecurity
health
Health IT
IBM
Watson
ethics
Defense Department
bots